

时序数据的异常检测可视化综述

1 介绍

时序数据被定义为一系列基于一个准确时间测量的结果，时间间隔通常是规律的[1]。例如按照一定时间间隔统计到的排名数据，实时检测的传感器数据，社交网络中每天的转发回复数据。

对于时序数据的分析在今天越来越广泛的应用在科学，工程，和商业领域，可视化帮助人们利用感知减少认知负荷进而理解数据[2]。长期以来，可视化也已经成功的被应用在对于时序数据的分析中来[3]。例如社交媒体[4]，城市数据[5]，电子交易[6]，时序排名[7]。在不同领域的时序数据中发现重要的特征和趋势的日益增长的需求刺激了许多可视交互探索工具的发展[8]：Line Graph Explore[9]，LiveRAC[2]，SignalLens[10]和 Data Vases[11]等。

时序数据的可视分析任务中，包括特征提取[14]，相关性分析和聚类[7]，模式识别[9]，异常检测[10]等。而异常检测在不同的研究领域都是一个重要的问题，异常检测表示发现数据中不符合预期行为的模式[12]。异常检测的目的是找到某些观察结果，它与其他的观察结果有很大的偏差，以至于引起人们怀疑它是由不同的机制产生的[17]。对应到不同的领域中，网络安全中的异常表示网络设备异常或者可疑的网络状态[13]。情感分析中的异常表示一组数据中反常的观点，情绪模式，或者产生这些模式的特殊时间[16]。社交媒体中的异常可以是反常的行为，例如识别网络机器人[20]，反常的传播过程，例如谣言的传播[19]。这些异常信息或模式的产生原因，可能是会影响日常生活，社会稳定的因素，例如电脑侵入，社交机器人，道路拥堵状况等。提早发现识别这些异常有助于及时找到产生原因和实际状况，从而进一步分析或解决问题。

异常检测已经有许多成熟的方法，而且在机器学习领域也引起了广泛的关注[12]，包括有监督[21]和无监督的异常检测方法[22]。自动化的学习算法通常基于这样的假设，即有充足的训练数据可用，同时这些数据理应是正常的行为，否则，正常的学习模型不能把新的观测结果按照异常来进行分类，很有可能新的观测数据是不常见的正常事件[25]，但当涉及到人工标注数据的问题时，往往需要大量的数据，费事费力，难以获取，同时又十分依赖于主观认为的判断，这些极大地影响了最后的分析结果质量[20]。与此同时，如何在自然数据中定义其中正常或异常的行为也是十分困难的[23]。此时人类的经验和知

识在异常检测方便是非常有价值，它可以用来更新改进模型，已经进一步的用作对异常检测过程的指导。而如今的大数据时代，面对数据维度多，数据尺寸大的场景，数据可视化具有强有力的功能和巨大的价值，可视化在其中可以更好的帮助人来分析理解数据和其中的行为，模式等。计算方法与人的经验和背景知识以及交互式可视化的灵活思维想结合，帮助人们发现从未想到的异常，减少人的劳动，提高异常的检测能力

2 挑战

异常检测的挑战在 Chandola V. [12]等人的综述中，被总结的很全面。

- a. 如何定义异常。正常和异常的界限往往是难以区分的，特别是当界限被规定后，在界限附近的异常观测，很容易把正常当做异常，亦或是把正常当做异常，例如一些**设定阈值的异常检测方法**，很难处理此种问题，需要其他的信息进行辅助判断。
- b. 有些**异常的行为**通常是人为的恶意操控，会模仿真实的行为，让异常的现象观测起来和正常现象一样，让异常检测的任务变得十分困难。例如社交网络中机器人[20]回复，它会模仿真实人类的语气，时间频率，以假乱真。
- c. 随着发展和进步，许多领域的正常行为也在与时俱进，其概念可能在未来会失效。而且很多数据集都是复杂和动态的，例如传感器数据[26][25]，网络安全数据等，这些挑战在[19]中都有提到。而在复杂多变时效性很高的场景中，需要人的监督和判断来进行异常的检测，分析，理解。
- d. 领域间的技术很难被应用于另一个领域，不同领域间的实际情况不一样，有些异常的产生在其它领域可能是正常的情况。
- e. 用于训练确认异常的模型的标记数据的获取以及可用性是很大的问题。
- f. 通常数据中包含的噪音和异常往往很相似，如何去区分和去除也是面临的问题。

数据

2.1 数据种类

2.2 数据属性

3 异常分类

不同的环境和实际情况中的时序数据，会包含许多的信息和属性，时序数据的异常情况便出现在附属在时间维度的信息上，例如社交网络中信息转发回

复的会话网络[19]，动态图的网络演变等数据和场景中包含了拓扑结构。例如网络数据中流量节点的类型，社交媒体上个人的信息[19]，注册时间等信息可以视为时序数据中的属性。异常检测的任务可以根据，拓扑结构和属性上的异常来进行分析。下面将从三个角度去对已有的工作进行分类，针对属性上的，拓扑结构上，和混合情况下的时序数据异常检测工作。例如时空数据中的地理信息，传感器网络的传输顺序，网络数据中的节点类型，动态图中的拓扑结构。

3.1 属性

Dennis Thom[15]等人的工作中根据信息的内容和发送的地理位置等属性进行聚类，形成标签云，用来可视分析时空数据中异常情况，例如地震，骚动等。Schreck T[18]等人的工作则是通和历史数据的内容频率进行对比，在地理视图上进行异常信息的高亮。[24]用于网络流量异常进行检测，对不同时间内的不同类型流量进行可视化，投影到六边形视图上。例如 DoS 攻击，探测攻击。[35]对运营数据中的异常流程通过在弦图中可视化每两个实例之间的关系，设定阈值已经通过判定关系之间的交叉和时序上的变化，进行异常如诈骗的检测。

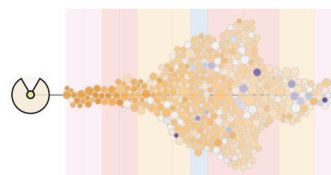
3.2 拓扑

原始数据中存在的拓扑结构，或者把数据抽象成拓扑结构

SAVE[26]提出了一个用于检测传感器数据异常变化的时序拓展模型(TEM)，在拓扑视图中，把传感器节点置于环形布局上，用颜色表示时间的先后顺序，可以观测传感器时序上的拓扑的变化，进而发现异常。设计维度相关性视图用于检测传感器数据的维度在时序上的变化的相关性的异常。

PolicyVis[27]系统用于防火墙安全策略的可视监测，巧妙的把安全策略的规则命令的逻辑顺序转化成拓扑结构，不同的异常对应不同的拓扑结构，按照时间和流量等维度绘制不同规则代表的矩形，矩形间的重合所带来的阴影代表了不同的情况，其中就包含如策略冲突等异常情况。

3.3 混合



Twitter 上消息的回复转发构成了会话网络，FluxFlow[19]系统用于对会话传播过程进行异常检测，探索，解释，它的分析模块结合了多种机器学习算法组来探索异常转发的特征，例如用户注册时间，朋友数量等属性上的特征。设计了一个时序上的包裹圆用于展示原始信息如何随着时间的推移在用户之间传播的可视化视图，并通过 OCCRF 模型[28]来计算其中每一个圆（用户）的异常分数，大小表示用户的重要性。结合拓扑上和属性上的特征进行异常检测。TargetVue[20]用于检测社交网络中具有异常行为的用户，系统最初为每个用户帐户提取一组行为特征，并使用异常检测算法来识别特征空间中的可疑用户。最可疑的用户用两种类型的图标进行可视化，一种是展示他们的通信行为（即发布消息，转发消息）；另一种是展示与之相应行为的特征，这些图标遵循类似的设计方案。

Qi Liao[34]等人的工作中研究管理网络数据中动态的异常问题，如社团中成员的关系，属性的变化。

异常检测类型	工作	数据
属性	[15] [18]	社交媒体，时空数据
	[24]	网络流量
拓扑	[26]	传感器
	[27]	防火墙策略
混合	[19]	社交媒体
	[34]	管理网络

4 异常检测方法(方法-可视化工具-数据源)

异常检测方法	领域	工作	数据源		可视化方法
聚类	社交媒体	[15]	Twitter		地理地图，标签云
对比	社交媒体	[18]	Twitter	对比信息内容和频率和历史对比	地理地图

无监督机器学习, One-Class Conditional Random Fields Model	社交媒体	[19]	Twitter Facebook		流图, 图标,
Self Organizing Maps (SOM),					
对比	网络攻击	[24]	the Darpa 1999		3D 图
聚类,GMM	传感器数据	[25]			3D, 地理地图, 柱状图, 散点图, 平行坐标图
对比, 无监督学习		[20]	Twitter		高位投影, Glyph, 拓扑图, 热力图
对比	传感器数据	[26]	GreenOrbs 森林传感器网络系统	SAVE	拓扑图, 折线图
	防火墙安全策略	[27]			
	管理网络	[34]			拓扑图
对比	运营数据	[35]			弦图

5 异常方法分类

5.1 特征统计

Celenk M.[29]基于短期的网络特征和平均时间熵的观测结果，对异常网络流量设计了 FLD 图，用于可视分析网络异常中的统计特征。Z-Glyph[23]被设计用于检测多元数据中的离群点，可以配合 small multiples 进行时序上异常检测。同样检测离群点的方还有，SOM[32]方法，用平行坐标轴[31]的方法，盒须图的相关方法[33]

2004 年[38]PCA 第一次被提出用于流量异常的检测，之后 Brauckhoff D 等人[37]提出了一个解决方案用于处理数据中时间相关性问题的，让 PCA 可以更好的应用于异常检测。还有基于直方图[40]，最大熵估计[39]等方法

6 总结和未来发展

本文对于时序数据的异常检测进行了综述，并针对现有的工作，对时序数据的异常类型和可视化方法分别进行了分类。异常检测需要先验知识，即人的判断来进行分析，可视化可以更好的对数据进行抽象，表达，发现自动算法不能判别的异常情况。

近年来，随着可视化方法在时序数据异常检测上的不断应用，逐步展现出可视化的巨大优势，例如社交领域[36]，现有的自动化方法的内在局限性，通过可视分析来检测异常的用户行为是一个十分有前途的方向，许多视觉分析系统通过用户专业的知识和经验，但是其中的挑战，例如不同领域的迁移，随时代变化的异常的判定标准和类型等，都将成为该研究方向在未来需要攻克的问题。

7 参考文献

- [1] Bojan V C, Raducu I G, Pop F, et al. Cloud-based service for time series analysis and visualisation in Farm Management System[C]//Intelligent Computer Communication and Processing (ICCP), 2015 IEEE International Conference on. IEEE, 2015: 425-432.
- [2] McLachlan P, Munzner T, Koutsofios E, et al. LiveRAC: interactive visual exploration of system management time-series data[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2008: 1483-1492.

- [3] Weber M, Alexa M, Müller W. Visualizing time-series on spirals[C]//Infovis. 2001, 1: 7-14.
- [4] Kumar P, Sinha A. Real-time analysis and visualization of online social media dynamics[C]//Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on. IEEE, 2016: 362-367.
- [5] Chen W, Huang Z, Wu F, et al. VAUD: A Visual Analysis Approach for Exploring Spatio-Temporal Urban Data[J]. IEEE Transactions on Visualization & Computer Graphics, 2017 (1): 1-1.
- [6] Xie C, Chen W, Huang X, et al. Vaet: A visual analytics approach for e-transactions time-series[J]. IEEE transactions on visualization and computer graphics, 2014, 20(12): 1743-1752.
- [7] Xia J, Hou Y, Chen Y V, et al. Visualizing Rank Time Series of Wikipedia Top-Viewed Pages[J]. IEEE Computer Graphics and Applications, 2017, 37(2): 42-53.
- [8] Cho M, Kim B, Bae H J, et al. Stroscope: Multi-scale visualization of irregularly measured time-series data[J]. IEEE transactions on visualization and computer graphics, 2014, 20(5): 808-821.
- [9] Kincaid R, Lam H. Line graph explorer: scalable display of line graphs using focus+ context[C]//Proceedings of the working conference on Advanced visual interfaces. ACM, 2006: 404-411.
- [10] Kincaid R. Signallens: Focus+ context applied to electronic time series[J]. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6): 900-907.
- [11] Thakur S, Rhyne T M. Data vases: 2d and 3d plots for visualizing multiple time series[J]. Advances in Visual Computing, 2009: 929-938.
- [12] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey[J]. ACM computing surveys (CSUR), 2009, 41(3): 15.
- [13] Pearlman J, Rheingans P. Visualizing network security events using compound glyphs from a service-oriented perspective[M]//VizSEC 2007. Springer, Berlin, Heidelberg, 2008: 131-146.
- [14] Alonso O, Khandelwal K. Kondenser: Exploration and visualization of archived social media[C]//Data Engineering (ICDE), 2014 IEEE 30th International Conference on. IEEE, 2014: 1202-1205.

- [15] Thom D, Bosch H, Koch S, et al. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages[C]//Pacific visualization symposium (PacificVis), 2012 IEEE. IEEE, 2012: 41-48.
- [16] Wang Z, Joo V, Tong C, et al. Anomaly Detection through Enhanced Sentiment Analysis on Social Media Data[C]//Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on. IEEE, 2014: 917-922.
- [17] Hawkins D M. Identification of outliers[M]. London: Chapman and Hall, 1980.
- [18] Schreck T, Keim D. Visual analysis of social media data[J]. Computer, 2013, 46(5): 68-75.
- [19] Zhao J, Cao N, Wen Z, et al. # FluxFlow: Visual analysis of anomalous information spreading on social media[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 1773-1782.
- [20] Cao N, Shi C, Lin S, et al. Targetvue: Visual analysis of anomalous user behaviors in online communication systems[J]. IEEE transactions on visualization and computer graphics, 2016, 22(1): 280-289.
- [21] Steinwart I, Hush D, Scovel C. A classification framework for anomaly detection[J]. Journal of Machine Learning Research, 2005, 6(Feb): 211-232.
- [22] Eskin E, Arnold A, Prerau M, et al. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data[J]. Applications of data mining in computer security, 2002, 6: 77-102.
- [23] Cao N, Lin Y R, Gotz D, et al. Z-Glyph: Visualizing outliers in multivariate data[J]. Information Visualization, 2017: 1473871616686635.
- [24] Onut I V, Zhu B, Ghorbani A A. A novel visualization technique for network anomaly detection[C]//PST. 2004: 167-174.
- [25] Riveiro M, Falkman G, Ziemke T. Improving maritime anomaly detection and situation awareness through interactive visualization[C]//Information Fusion, 2008 11th International Conference on. IEEE, 2008: 1-8.
- [26] Shi L, Liao Q, He Y, et al. SAVE: Sensor anomaly visualization engine[C]//Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on. IEEE, 2011: 201-210.
- [27] Tran T, Al-Shaer E S, Boutaba R. PolicyVis: Firewall Security Policy Visualization and Inspection[C]//LISA. 2007, 7: 1-16.

- [28] Song Y, Wen Z, Lin C Y, et al. One-Class Conditional Random Fields for Sequential Anomaly Detection[C]//IJCAI. 2013: 1685-1691.
- [29] Celenk M, Conley T, Willis J, et al. Predictive network anomaly detection and visualization[J]. IEEE Transactions on Information Forensics and Security, 2010, 5(2): 288-299.
- [30] Kandogan E. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates[C]//Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001: 107-116.
- [31] Novotny M, Hauser H. Outlier-preserving focus+ context visualization in parallel coordinates[J]. IEEE Transactions on Visualization and Computer Graphics, 2006, 12(5): 893-900.
- [32] Munoz A, Muruzábal J. Self-organizing maps for outlier detection[J]. Neurocomputing, 1998, 18(1): 33-60.
- [33] Kampstra P. Beanplot: A boxplot alternative for visual comparison of distributions[J]. 2008.
- [34] Liao Q, Striegel A. Intelligent network management using graph differential anomaly visualization[C]//Network Operations and Management Symposium (NOMS), 2012 IEEE. IEEE, 2012: 1008-1014.
- [35] Hao M C, Keim D A, Dayal U, et al. Business process impact visualization and anomaly detection[J]. Information Visualization, 2006, 5(1): 15-27.
- [36] Wu Y, Cao N, Gotz D, et al. A survey on visual analytics of social media data[J]. IEEE Transactions on Multimedia, 2016, 18(11): 2135-2148.
- [37] Brauckhoff D, Salamatian K, May M. Applying PCA for traffic anomaly detection: Problems and solutions[C]//INFOCOM 2009, IEEE. IEEE, 2009: 2866-2870.
- [38] Lakhina A, Crovella M, Diot C. Diagnosing network-wide traffic anomalies[C]//ACM SIGCOMM Computer Communication Review. ACM, 2004, 34(4): 219-230.
- [39] Gu Y, McCallum A, Towsley D. Detecting anomalies in network traffic using maximum entropy estimation[C]//Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement. USENIX Association, 2005: 32-32.
- [40] Kind A, Stoecklin M P, Dimitropoulos X. Histogram-based traffic anomaly detection[J]. IEEE Transactions on Network and Service Management, 2009, 6(2).
- [41]